Forecasting Individual Demand in Cellular Networks

Guangshuo Chen[1,3], Sahar Hoteit[2], Aline Carneiro Viana[3], Marco Fiore[4] and Carlos Sarraute[5]

[1]*École Polytechnique, Université Paris-Saclay, France*
[2]*L2S, Université Paris Sud-CNRS-CentraleSupélec, Université Paris-Saclay, France*
[3]*INRIA, Université Paris-Saclay, France* [4]*CNR - IEIIT, Italy* [5]*Grandata Labs, USA*

We leverage two large-scale real-world datasets to provide the pioneer results on the limits of predictability of per-user mobile data traffic demands over time and space. Using information theory tools, we measure the maximum predictability that any algorithm has potential to achieve. We first focus on the predictability of mobile data traffic consumption patterns in isolation. Our results show that it is theoretically possible to anticipate individual demands with a typical accuracy of 85%. Then, we analyze the joint predictability of mobile data traffic demands and mobility patterns. Their correlation that we find leads to a higher theoretical potential performance in joint prediction. Besides, we propose a novel practice to evaluate the spatiotemporal correlation of per-user mobile data traffic.

Keywords: Fundamental limits; user mobility; user data traffic; call detail records; performance analysis

1 Introduction

The quantitative understanding of human behaviors (*e.g.*, user's whereabouts or mobile data traffic demands) has recently emerged as a central question in multi-disciplinary research [1]. In the context of forecasting human behaviors, any practical technique's performance is bounded theoretically by *predictability* which measures to what degree a specific behavior can be foreseen. In this paper, we analyze the predictability of mobile data traffic from the viewpoint of individual users, in terms of their consumed data volumes and whereabouts. Our study allows answering an important question: *to what degree is the individual consumption of mobile data traffic predictable?* To the best of our knowledge, there is no analysis of *(i)* how per-user regularity of mobile data traffic is translated into actual predictability, or *(ii)* the associated impacts to predictability brought by jointly considering users' visited locations. We refer the reader to [2] for a more elaborate version of the literature review.

In this paper, we evaluate the predictability by studying per-user variations of mobile data traffic over time and space and investigating their predictability limits by using tools from information theory. Based on two large-scale real-world datasets, our results reveal a promising upper bound (85%) to the performance of practical algorithms that forecast future mobile data traffic volumes from a user's historical usage. Then, our pioneer investigation reveals a strong correlation between mobility and mobile service usage. Such correlation, on the power of jointly forecasting when, where, and how much mobile data traffic is generated by individual users, leads to a performance gain (10%) in terms of the joint predictability. Finally, our discussion about the cause of the high (joint) predictability sheds light on the design of predicting algorithms.

2 Data overview

Our study is based on the behavioral data of footprints and mobile data traffic demands of 45K anonymous mobile network subscribers during 92 consecutive days in 2014. The users come from the capital city of a Latin American country. Each user has (1) *call detail records*, *i.e.*, time-stamped and geo-referenced logs attached to all his voice calls, and (2) *session records*, *i.e.*, time-stamped logs of Internet data sessions with

data volumes. The users meet the following criteria: *(i)* they have visited at least 2 locations; *(ii)* they have footprints in at least 20% of the observing hours; *(iii)* they establish data sessions in at least 73 days (80% of the observing period). The criteria ensure statistical significance of our analysis.

We compute for each user u two representative discretized time series (1) of data traffic volumes, marked as $v_1^T(u)$, and (2) of locations, marked as $\ell_1^T(u)$, both of which cover the same $T = 24$ hours/day $\times$ 92 days $=$ 2208 time slots. For the former, data session records are aggregated and quantized in a straight-forward way. In a time series $v_1^T(u) = \{v_u^1, \cdots, v_u^T\}$, each discretized volume v_u^i is marked by one of the eight quantizations, *i.e.*, 0 (*idle*), $(1, 10)$, $(10, 10^2)$, ..., $(10^6, 10^7)$ in kilobytes, representing the aggregated mobile data traffic of the corresponding time slot. For the latter, in the i-th time slot, ℓ_u^i of a time series $\ell_1^T(u) = \{\ell_u^1, \cdots, \ell_u^T\}$ is the representative location (which has the most frequent appearance) of that time slot. Note that the original location source (from voice calls) cannot offer complete mobility information. Thus, we apply the `stop-by-spothome` approach [3] on call detail records in advance, to enhance the temporal coverage of locations, particularly overnight, without affecting the localization precision. For more details of the data and preprocessing, we refer the reader to the full version [2].

3 Predictability of mobile data traffic

We first study the predictability of mobile data traffic generated by individual subscribers, particularly, on the forecast of mobile data traffic volumes in isolation. For each user u, we consider that his discrete volumes come from a random process $\mathcal{V} = \{V_t\}$ and compute the entropy rate $H_u(\mathcal{V}) = \lim_{T\to+\infty} \frac{1}{T} \sum_{t=1}^{T} H(V_t|V_{t-1}, \cdots, V_1)$ that represents the average uncertainty of discrete volumes at each time slot given preceding volumes as a prior knowledge. We derive three entropy rate variants from different models. *(i)* The *temporal-uncorrelated entropy* rate is formulated as $H_u^{\text{unc}}(\mathcal{V}) \equiv -\sum_{v\in v_1^T(u)} P(v) \log P(v)$ where the user's traffic follows a heterogeneous and time-independent model. This entropy rate characterizes the heterogeneity of a mobile demand model that has no temporal correlations, hence its name. *(ii)* The *nonzero-temporal-uncorrelated entropy* rate is based on the same model of $H_u^{\text{unc}}(\mathcal{V})$, but it is limited to those cases when the user is not idle. Formally, it is $H_u^{\text{n0}}(\mathcal{V}) \equiv -\sum_{v\in v_1^T(u)/\{0\}} P(v|v \neq 0) \log P(v|v \neq 0)$. It captures the heterogeneity of data traffic volumes exchanged during active hours only, yet still ignoring temporal correlations. *(iii)* The *actual entropy* rate $H_u(\mathcal{V})$ depends not only on the frequency of appearance of each discretized traffic volume but also on the order in which they appear, capturing the temporal order presented in a subscriber's data traffic usage pattern. Since we only have a finite time series $v_1^T(u)$, we employ an estimator based on Lempel-Ziv compression [4] to compute $H_u(\mathcal{V})$. It is worth noting that we favor the binary logarithm so that the unit of all entropy rate variants is *bit*.

These entropy rate variants are computed as a prepositive step of measuring predictability. Intuitively, entropy (rate) and predictability are negatively correlated variables: a behavior with low (or high) uncertainty is highly (or little) predictable. Mathematically, given an entropy rate variant H, its predictability Π satisfies $\Pi \leq \Phi^{-1}(H)$ where $\Phi(x) \equiv (1-x)log(N-1) - x\log x - (1-x)\log(1-x)$ and $\Phi^{-1}(x)$ is its inverse function [5]. In our context, an upper bound is an estimation of the maximum achievable accuracy in the prediction of mobile data traffic demands given a particular model. Hence, three upper bounds of the predictability, *i.e.*, $\Pi_u^{\text{max}}(\mathcal{V})$, $\Pi_u^{\text{unc}}(\mathcal{V})$, and $\Pi_u^{\text{n0}}(\mathcal{V})$, are calculated from the corresponding entropy variants.

We portray the PDF (probability density function) of all the entropy rate variants in Fig. 1(a). Let us start with the temporal-uncorrelated entropy $H_u^{\text{unc}}(\mathcal{V})$. It has a high peak at $2^{H_u^{\text{unc}}(\mathcal{V})} = 2^{1.63} \approx 3$, indicating that each user tends to generate data traffic that is described by just three quantization levels (although there are eight) when we only focus on the overall probability of each level's appearance. Interestingly, idle time intervals do not bias such regularity. Indeed, the PDF of $H_u^{\text{n0}}(\mathcal{V})$ overlaps well to that of $H_u^{\text{unc}}(\mathcal{V})$, suggesting that the considerations above also hold when only time intervals with data sessions are considered. However, our main result is the significant shift presented by the PDF of $H_u(\mathcal{V})$, which peaks at 0.97. When taking the temporal ordering of data sessions into account, one can reduce the uncertainty to just two quantization levels.

The distributions in Fig. 1(b) confirm the findings above and provide upper numerical bounds to the predictability of per-user mobile data traffic demand. We observe that $\Pi_u^{\text{unc}}(\mathcal{V})$ and $\Pi_u^{\text{n0}}(\mathcal{V})$ at 0.69 and

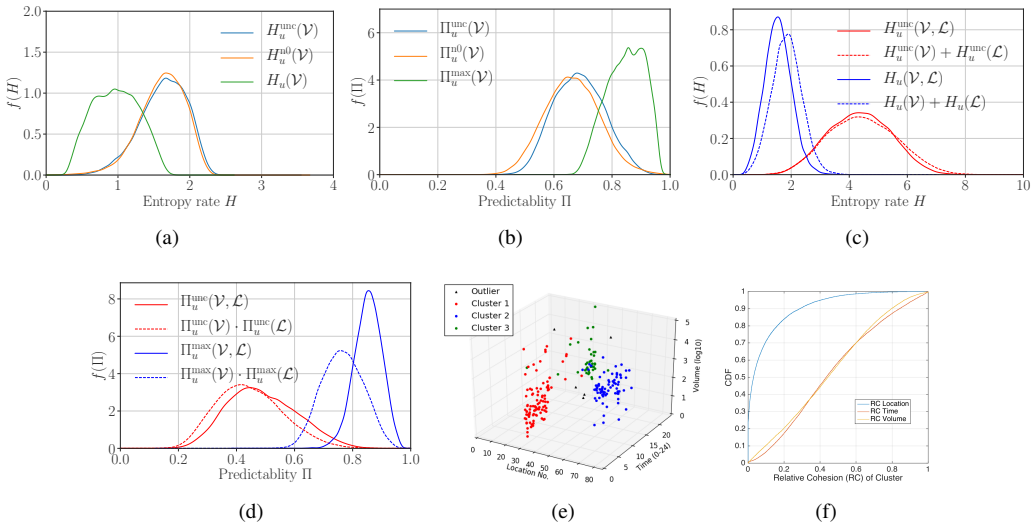

Figure 1: (Best viewed in colors) (a) Distributions of the per-user entropy rate variants of discrete data traffic volumes in isolation. (b) Corresponding distributions of upper bounds on the predictability of data traffic volumes in isolation. (c) Distributions of the different flavors of joint entropy rate variants. (d) Distributions of the corresponding joint predictability upper bounds. (e) An example of mapping a user's data sessions into a three-dimensional space. (f) CDF of each cluster's relative cohesion $RC^{(*)}$ on the three dimensions.

0.66, respectively, which means that a relatively good predictability can be possibly achieved even if we do not consider temporal regularity in prediction. More importantly, $\Pi_u^{\max}(\mathcal{V})$ indicates that the demand of a subscriber can be possibly predicted within 85% accuracy on average. It means that in only 15% of the time does a user generate data traffic volumes in a manner which appears to be random, but in the remaining 85% of the time, we could hope to predict his volume accurately. This result proves, for the first time, that *data traffic volumes which subscribers generate via their mobile devices are highly predictable*.

4 Joint predictability of traffic and mobility

We further study the joint predictability of future mobile data traffic volumes and visited locations on a per-user basis. We investigate how predictable is the combination of *how much* traffic is generated by a mobile phone user and *where* this happens on each time slot. First, leveraging $\ell_1^T(u)$, we measure two entropy rate variants on mobility of each user u, *i.e.*, the temporal-uncorrelated entropy rate $H_u^{\mathrm{unc}}(\mathcal{L})$ and the actual entropy rate $H_u(\mathcal{L})$, as well as their corresponding predictability upper bounds $\Pi_u^{\mathrm{unc}}(\mathcal{L})$ and $\Pi_u^{\max}(\mathcal{L})$. Then, combining $\ell_1^T(u)$ and $v_1^T(u)$, we compute two joint entropy rate variants that consider volumes and locations of each user together. The *temporal-uncorrelated entropy* rate $H_u^{\mathrm{unc}}(\mathcal{V},\mathcal{L}) \equiv -\sum_{v\in v_1^T(u),\ell\in\ell_1^T(u)} P(v,\ell)\log P(v,\ell)$ determines the joint heterogeneity of a user's locations and data traffic volumes. The *joint actual entropy rate* $H_u(\mathcal{V},\mathcal{L})$ is defined as the actual entropy rate of the joint stationary process $\{(V_t,L_t)\}$. It expresses the combined uncertainty of a user's locations and data traffic volumes on each time slot, considering his previous history of movements and mobile service usage. Also, the corresponding predictability upper bounds $\Pi_u^{\mathrm{unc}}(\mathcal{V},\mathcal{L})$ and $\Pi_u^{\max}(\mathcal{V},\mathcal{L})$ are calculated.

Figs. 1(c) and 1(d) concern our actual measures of interest with respect to the uncertainty and predictability of mobile data traffic and mobility. A first interesting remark is that, $H_u^{\mathrm{unc}}(\mathcal{V},\mathcal{L})$ and $H_u^{\mathrm{unc}}(\mathcal{V})+H_u^{\mathrm{unc}}(\mathcal{L})$, and consequently $\Pi_u^{\mathrm{unc}}(\mathcal{V},\mathcal{L})$ and $\Pi_u^{\mathrm{unc}}(\mathcal{V})\cdot\Pi_u^{\mathrm{unc}}(\mathcal{L})$, are nearly indistinguishable. Instead, $H_u(\mathcal{V},\mathcal{L})$ and $H_u(\mathcal{V})+H_u(\mathcal{L})$, and consequently $\Pi_u^{\max}(\mathcal{V},\mathcal{L})$ and $\Pi_u^{\max}(\mathcal{L})\cdot\Pi_u^{\max}(\mathcal{V})$, show significant differences. Accordingly, there exists some correlation between mobility and data traffic consumption processes. Such correlation mainly emerges when considering – and it is thus driven by – the temporal ordering of events. As observed in Fig. 1(d), a joint prediction of the next consumed data traffic amount and location where this occurs, from the user's previous actions, can yield a better accuracy than forecasting the two

separately, since the shift between $\Pi_u^{\max}(\mathcal{L}) \cdot \Pi_u^{\max}(\mathcal{V})$ and $\Pi_u^{\max}(\mathcal{V},\mathcal{L})$ is 10%. Therefore, our main conclusion is that *it is possible to anticipate how much mobile data traffic (as an order of magnitude) will be consumed by a given user and where this will occur in a very effective manner (*i.e.*, with an* 88% *accuracy on average), by knowing the historical activities.*

5 Discussion

To understand the cause of the high (joint) predictability, we map each user's data sessions into a three-dimensional space of *location*, *time*, and *volume*, so as to have an intuitive spatiotemporal representation. Each session becomes then a point $p(l,t,v)$ into this space. Note that we express l as the linear ordering of the corresponding bidimentional locations, as returned by the density-based Optics cluster algorithm that places spatially close bidimentional locations as neighbors in the ordering l. Time t is expressed by hours with decimals from 0 to 24, where the date is ignored. Volume v is the magnitude of the traffic volume, *i.e.*, $log_{10}(\cdot)$. This mapping reveal how a user generates mobile Internet sessions. Fig. 1(e), most sessions are aggregated on two major locations (30 and 60), probably mapping to home and working place according to their time of visits. Sessions containing data traffic over 10MB mainly occur at the location 30 during nighttime. To quantitatively investigate the 3D space representation of $p(l,t,v)$ points, we use DBScan to cluster each user's sessions in the three-dimensional space. For the clustering, a weighted euclidean distance is measured between every two points $p_1(l_1,t_1,v_1)$ and $p_2(l_2,t_2,v_2)$, where the distance of each dimension is computed as follows: *(i)* $dist^{(location)}(p_1,p_2) = \omega_l|\mathbf{l_1} - \mathbf{l_2}|_{geo}$ in kilometers; *(ii)* $dist^{(time)}(p_1,p_2) = \omega_t|t_1 - t_2|$ in hours; *(iii)* $dist^{(volume)}(p_1,p_2) = \omega_v|v_1 - v_2|$ ($|log_{10}\frac{Vol_1}{Vol_2}|$). Each distance is applied to 99% percentile normalization. For each cluster shown in Fig. 1(e), we then use the *relative cohesion* (RC) to quantify the contribution of each dimension to a given cluster as $RC^{(*)} = \frac{\sum_{p\in C} dist^{(*)}(p,c)^2}{\sum_{p\in C} dist(p,c)^2}$, where C and c respectively represent the cluster and its centroid $c = (l_{centroid}, t_{mean}, v_{mean})$. The RCs of the three dimensions satisfy $RC^{(loc)} + RC^{(time)} + RC^{(vol)} = 1$, where $0 < RC^{(*)} < 1$. Hence, if a cluster's RC in one dimension is significantly smaller than the other two dimensions, this dimension is contributing the most to creating the cluster.

Fig. 1(f) shows the distributions of RCs along the three dimensions for all users. The most striking behavior is the much lower RC in space than time or traffic volume: *i.e.*, *"where a user is"* drives the creation of majority clusters: The location of a mobile user has a high probability to trigger some routine service consumption activity. Hence, anticipating the future location of a user should be the first target of a solution aiming at predicting mobile user activity. However, we also observe that locations alone do not explain all clusters. A non-negligible fraction of clusters showing high RC in space and low RC in time and traffic volume are also present in several cases. We conclude that the three dimensions are complementary, and though different weights, they are all important for an accurate prediction of mobile users' behaviors. This is consistent with – and explains – our results on the high joint predictability of temporally correlated visited locations and consumed traffic. Consequently, our results indicate that there is a large space for predicting mobile data traffic and adapting network optimizing solutions based on the latter, such as for load balancing.

References

[1] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale Mobile Traffic Analysis: a Survey," *IEEE Communications Surveys & Tutorials*, vol. PP, no. 99, pp. 1–1, 2015.

[2] G. Chen, S. Hoteit, A. Carneiro Viana, M. Fiore, and C. Sarraute, "Spatio-Temporal Predictability of Cellular Data Traffic," Research Report RT-0483, INRIA Saclay - Ile-de-France, Jan. 2017.

[3] S. Hoteit, G. Chen, A. Viana, and M. Fiore, "Filling the gaps: On the completion of sparse call detail records for mobility analysis," in *ACM Chants*, 2016.

[4] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 6, pp. 414–427, Sept. 1996.

[5] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, 1994.